ECON3150/4150 – Introductory Econometrics – Open book, home exam

Exam date and time: Monday, May 25, 2020 from 09.00 – 14.00 (five hours)

Language: The examination text is given in English. You may submit your response in Norwegian, Swedish, Danish or English.

Guidelines: You should upload your text in pdf format - **one pdf per question.** You can scroll back and forth in the problem set.

You should familiarize yourself with the rules that apply to the use of sources and citations.

The answers to your exams are not expected to meet the formal requirements for references and citations in the spring 2020 exam. However, you should make references by indicating the source in the text. Creating a bibliography is not required. Whether you choose to do so, or not has no impact on your grade. The purpose of a reference is that the examiner should be able to look up the source him/herself, either to read it or to evaluate your interpretation. If you are referring to a limited part of the source, the reference should indicate which part of the source you refer to by using page numbers. If you are quoting directly from a source, follow the normal citation practice – with quotations marks and references to the source.

The exam lasts for only five hours. We recommend that you use the available time to work on the problem set, as well as allocate time to scan attachments with graphs and/or equations.

The problem set: The problem set consists of two questions, with several sub-questions. They count as indicated. Start by reading through the whole exam, and make sure that you allocate time to answering problems you find easy.

Note: You can resize the question by clicking on the three dots on the right, hold and pull to the right. Similarly for the three dots at the bottom, click, hold and pull down. Then the text will be larger. You can also click on the link on the right side of the screen and download the problem text to your own machine.

Digital hand drawings/graphs/equations: You will find information about options for hand drawings on this website:
https://www.uio.no/english/studies/examinations/submissions/options-for-hand-drawings.html

Submission in Inspera

- Read more about exam and submission in Inspera. https://www.uio.no/english/studies/examinations/submissions/.
- When your answer is uploaded, you will see that the exam is uploaded and saved.

- To submit your answer, please see https://www.uio.no/english/studies/examinations/submissions/submit_answer/. You can either choose the “submit now” or the “Automatic submission”.
- You can make changes in your exam until the deadline.
- You will find the answer under Archives.

Do you need technical support, or do you have any questions during the exam?

Please send an e-mail, titled “ECON3150/4150” To hjemmeeksamen@sv.uio.no from your university email.

Grading: The grades given: A-F, with A as the best and E as the weakest passing grade. F is fail.

Grades are given: Monday, June 15, 2020.

Question 1

A researcher wants to investigate if the number of hours that children go to school during a year affects test scores. She has a panel data set with information on 150 regions for the years 2000-2010. The dependent variable $testscore_{it}$ is the average test score (in points) obtained by students in region i in year t. The explanatory variable $hours\ in\ school_{it}$ is the number of hours that students spent in school in region i in year t.

a) The researcher decides to estimate the following regression model by OLS

$$testscore_{it} = \beta_0 + \beta_1 \cdot hours\ in\ school_{it} + u_{it} \tag{1}$$

She obtains the following estimation results

```
model1 <- lm( testscore ~ hours_in_school, data = data)
coeftest(model1,vcovHC(model1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     -16.349402   2.494307 -6.5547 7.437e-11
## hours_in_school   0.432520   0.014356 [redacted]
##
```

Give an interpretation, in words, of the estimated coefficient $\widehat{\beta}_1$.

b) Is the coefficient on $hours\ in\ school_{it}$ significantly different from zero at a 1 percent significance level?

c) The researcher decides to take the logarithm of test scores as dependent variable and estimates the following regression model by OLS

$$ln(testscore_{it}) = \pi_0 + \pi_1 \cdot hours\ in\ school_{it} + \varepsilon_{it} \tag{2}$$

She obtains the following estimation results

```
model2 <- lm( ln_testscore ~ hours_in_school, data = data)
coeftest(model2,vcovHC(model2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     2.55937207 0.05911901  43.292 < 2.2e-16
## hours_in_school 0.00845617 0.00033237  25.442 < 2.2e-16
```

Give an interpretation, in words, of the estimated coefficient $\widehat{\pi}_1$.

d) Compute a 95 percent confidence interval around π_1.

e) Do you think that the OLS estimator of π_1 is an unbiased estimator of the causal effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$? Explain why or why not.

f) The researcher decides to use an instrumental variable approach to estimate the causal effect of hours spent in school on average test scores. In 2005 there was a pandemic and all schools were closed for part of the year. She decides to create a binary variable $pandemic_t$ which equals one for all regions in 2005 and zero otherwise. She estimates the following first stage regression model by OLS

$$hours\ in\ school_{it} = \delta_0 + \delta_1 \cdot pandemic_t + \epsilon_{it} \tag{3}$$

and obtains the following estimation results

```
first_stage <- lm( hours_in_school ~ pandemic, data = data)
coeftest(first_stage,vcovHC(first_stage, type = "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 173.72667    0.42246  411.23 < 2.2e-16 ***
## pandemic    -59.52667    1.29461 [redacted]
```

Do you think that the instrument relevance condition holds? Is $pandemic_t$ a weak instrument?

g) The researcher estimates the following regression model by OLS

$$ln(testscore_{it}) = \gamma_0 + \gamma_1 \cdot pandemic_t + \upsilon_{it}$$

and obtains the following estimation results.

```
reduced_form<- lm( ln_testscore ~ pandemic, data = data)
coeftest(reduced_form,vcovHC(reduced_form, type = "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  4.004575   0.008133 492.3865 < 2.2e-16 ***
## pandemic    -0.240918   0.035967  -6.6983 2.884e-11 ***
```

Use these results in combination with the first stage results from part f) to compute the instrumental variable estimate of the effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$.

h) Do you think that, when using $pandemic_t$ as an instrument to estimate the causal effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$, the instrument exogeneity condition holds? Explain why or why not.

i) Instead of using an instrumental variable approach the researcher decides to include region fixed effects. She estimates the following regression model

$$ln(testscore_{it}) = \theta_0 + \theta_1 \cdot hours\ in\ school_{it} + \eta_i + \nu_{it} \tag{4}$$

and obtains the following estimation results.

```
within <- plm(ln_testscore ~ hours_in_school, data = data,
          index = c("region_id"), model = "within")
class(within)
```

```
## [1] "plm"        "panelmodel"
```

```
coeftest(within,vcovHC(within, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## hours_in_school 0.00379138 0.00032263  11.751 < 2.2e-16
```

Compare these results to the results in part c) and explain whether the results differ and if so why.

j) The researcher thinks that estimating the following model by OLS

$$ln(testscore_{it}) - ln(testscore_{it-1}) = \theta_1 \cdot (hours\ in\ school_{it} - hours\ in\ school_{it-1}) + (\nu_{it} - \nu_{it-1}) \tag{5}$$

will give an identical estimate of the causal effect of $hours\ in\ school_{it}$ on $ln(testscore_{it})$ as the estimate shown in the R-output in part i). Is she right, explain why or why not.

k) The test is in English and in some regions students don't have English as their native language. The researchers thinks this might affect test scores and decides to include the binary variable $no\ english_i$ which equals one for regions where students don't have English as their native language and zero otherwise. She estimates the following regression model

$$ln(testscore_{it}) = \theta_0 + \theta_1 \cdot hours\ in\ school_{it} + \theta_2 \cdot no\ english_i + \eta_i + \omega_{it} \tag{6}$$

and obtains the following estimation results.

```
within2 <- plm(ln_testscore ~ hours_in_school + no_english,
          data = data, index = c("region_id"), model = "within")
class(within)
```

```
## [1] "plm"        "panelmodel"
```

```
coeftest(within2,vcovHC(within2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## hours_in_school 0.00379138 0.00032263  11.751 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explain why the R-output does not show an estimated coefficient on $no\ english_i$. Is it possible to estimate the coefficient on the variable $no\ english_i$ when region fixed effects are included in the regression model?

l) The researcher wants to control for omitted variables that are common across regions but that vary over time and decides to include year fixed effects. She estimates the following regression model

$$ln(testscore_{it}) = \theta_0 + \theta_1 \cdot hours\,in\,school_{it} + \eta_i + \tau_1 \cdot year2001 + ... + \tau_{10} \cdot year2010 + \mu_{it} \tag{7}$$

She wants to test whether the time fixed effects are jointly significantly different from zero and performs an F-test with the following results:

```r
linearHypothesis(within3, c("year2001", "year2002", "year2003", "year2004",
                            "year2005", "year2006", "year2007", "year2008",
                            "year2009","year2010" ),
                 test=c("F"), vcov = vcovHC(within3, type = "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## year2001 = 0
## year2002 = 0
## year2003 = 0
## year2004 = 0
## year2005 = 0
## year2006 = 0
## year2007 = 0
## year2008 = 0
## year2009 = 0
## year2010 = 0
##
## Model 1: restricted model
## Model 2: ln_testscore ~ hours_in_school + year
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1 [redacted]
## 2 [redacted] 1.4715 [redacted]
```

Are the time fixed effects jointly significantly different from zero at a 1 percent significance level?

Question 2

A business owner wants to know if bonus payments will increase the work effort of the employees. He asks his research department to set up an experiment in order to estimate the average causal effect of bonus payments on work effort. The research department randomly assigns 500 employees either to a treatment group or a control group. The 250 employees assigned to the treatment group receive a bonus if they meet the target, the 250 employees in the control group do not get a bonus if they meet the target. The experiment lasts for 3 months and at the end of the period the research department collects information on work effort. They construct a binary variable $effort_i$ which equals one if the worker exerted high effort during the 3 months and zero if the worker exerted low effort. The data set collected by the research department contains in addition a binary variable $bonus_i$ which equals one if the worker was assigned to the treatment group and zero if assigned to the control group and the variable $female_i$ which equals one for female employees and zero for male employees.

a) The research department decides to estimate the following regression model by OLS

$$effort_i = \beta_0 + \beta_1 \cdot bonus_i + u_i \tag{1}$$

and obtains the following estimation results

```
model1 <- lm( effort ~ bonus, data = data2)
coeftest(model1,vcovHC(model1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.440000   0.031457 13.9872  < 2e-16
## bonus       0.096000   0.044591  2.1529  0.03181
## ---
```

Give an interpretation, in words, of the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

b) The experiment started during the summer holiday and all 500 workers of the firm were randomly assigned to the treatment or control group. Part of these workers are students. Midway during the experiment, the summer holiday ends and all students quit their job and go back to school. These students are therefore not part of the data set collected by the research department. Do you think that the OLS estimator of β_1 in model (1) (estimated in part (a)) is a consistent estimator of the causal effect of bonus payments on the probability of exerting high effort? Explain why or why not.

c) The business owner wants to know if men and women respond differently to bonus payments. In order to answer this question the research department decides to estimate the following regression model by OLS

$$effort_i = \beta_0 + \beta_1 \cdot bonus_i + \beta_2 \cdot female_i + \beta_3(bonus_i \times female_i) + \epsilon_i \tag{2}$$

and obtains the following estimation results

```
model2 <- lm( effort ~ bonus + female + (female*bonus), data = data2)
coeftest(model2,vcovHC(model2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   0.338843   0.043202 [redacted]
## bonus         0.234800   0.061462
## female        0.196041   0.061729
## bonus:female -0.273816   0.088342
## ---
```

Give an interpretation, in words, of the estimated coefficient $\widehat{\beta}_3$. What is the estimated effect of bonus payments on the probability of exerting high effort for men and for women?

d) The business owner wants to know if the probability of exerting high effort differs significantly between men and women in absence of a bonus payment. Use the results of part c) and a 5 percent significance level to answer the question of the business owner.

e) The business owner wants to know if the probability of exerting high effort differs significantly between men and women in case workers receive a bonus payment if they meet the target. Explain how the research department can answer the question of the business owner.

f) The research department decides to estimate a logit model and they obtain the following estimation results

```
logit <- glm(effort ~ bonus + female + (female*bonus),
             family = binomial(link = "logit"),
             data = data2)

coeftest(logit,vcovHC(logit, type = "HC1"))
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -0.66845    0.19284 -3.4663 0.0005276
## bonus         0.96519    0.26294  3.6707 0.0002419
## female        0.80822    0.26191  3.0858 0.0020299
## bonus:female -1.12148    0.36589 -3.0651 0.0021763
## ---
```

What is the estimated effect of bonus payments on the probability of exerting high effort for men and for women?

g) Does the 99 percent confidence interval around the logit coefficient on the interaction term between $female_i$ and $bonus_i$ include the value zero?

h) The research department decides to estimate a probit model and they obtain the following estimation results

```
probit <- glm(effort ~ bonus + female + (female*bonus)
             family = binomial(link = "probit"),
             data = data2)

coeftest(probit,vcovHC(probit, type = "HC1"))
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -0.41562    0.11806 -3.5204 0.0004308
## bonus         0.60128    0.16238  3.7029 0.0002132
## female        0.50318    0.16201  3.1058 0.0018975
## bonus:female -0.69919    0.22752 -3.0731 0.0021182
## ---
```

What is the estimated effect of bonus payments on the probability of exerting high effort for men and for women?